Habitat-Lite: A GSC case study based on free text terms for environmental metadata

Lynette Hirschman[1], Cheryl Clark[1], K. Bretonnel Cohen[1], Scott Mardis[1], Joanne Luciano[1], Renzo Kottmann[2], James Cole[3], Victor Markowitz[4], Nikos Kyrpides[5], Dawn Field[6]

[1] Information Technology Center, The MITRE Corporation, 202 Burlington Rd., Bedford, MA 01730, USA

[2] Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, 28359 Bremen, Germany

[3] Center For Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA

[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[5] Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA.

[6] NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, Oxfordshire OX1 3SR, UK

Corresponding author:

Lynette Hirschman

Information Technology Center

The MITRE Corporation

202 Burlington Rd.

Bedford, MA 01730

USA

781-271-7789

lynette@mitre.org

cclark@mitre.org

kbcohen@mitre.org

mardis@mitre.org

jluciano@mitre.org

rkottman@mpi-bremen.de

colej@msu.edu

vmmarkowitz@lbl.gov

NCKyrpides@lbl.gov

dfield@ceh.ac.uk

Keywords

Annotation, Controlled vocabulary, Ontology, Habitat, Metadata, Text mining, Curation, Environment

Abstract

There is an urgent need to capture metadata on the rapidly growing number of genomic, metagenomic and related sequences, such as 16S ribosomal genes. This need is a major focus within the Genomic Standards Consortium (GSC), and *Habitat* is a key metadata descriptor in the proposed "Minimum Information about a Genome Sequence" (MIGS) specification. The goal of the work described here is to provide a light-weight, easy-to-use (small) set of terms ("Habitat-Lite") that captures high-level information about habitat while preserving a mapping to the recently launched Environment Ontology (EnvO). Our motivation for building Habitat-Lite is to meet the needs of multiple users, such as annotators curating these data, database providers hosting the data, and biologists and bioinformaticians alike who need to search and employ such data in comparative analyses. Here, we report a case study based on semi-automated identification of terms from GenBank and GOLD. We estimate that the terms in the initial version of Habitat-Lite would provide useful labels for over 60% of the kinds of information found in the GenBank isolation_source field, and around 85% of the terms in the GOLD habitat field. We present a revised version of Habitat-Lite and invite the community's feedback on its further development in order to provide a minimum list of terms to capture high-level habitat information and to provide classification bins needed for future studies.

Introduction

This paper discusses the current status of an ongoing effort to create a minimum hierarchical controlled vocabulary for the capture of habitat and environmental metadata on genomics, metagenomics and 16S ribosomal sequences. This work has two goals. The short-term goal is to develop a light-weight controlled vocabulary (Habitat-Lite) to capture high-level habitat and environmental metadata in support of the Genomic Standards Consortium (GSC) Minimal Information about Genome/Metagenome Sequence (MIGS/MIMS) specification (Field et al. 2008; Field et al. 2008a). The longer-term goal is to develop a repeatable process for other types of metadata by identifying key terms based on usage in databases and the open literature. We will evaluate the coverage, utility, and usability of the key terms and refine the set of terms based on these measures. Additionally, we will develop tools to facilitate the capture of the metadata from free text fields.

This effort originated in the context of the development of the MIGS/MIMS checklist[1], and has also been discussed in the context of the newly established Environment Ontology (EnvO) project[2], as part of advocating the use of ontologies in capturing MIGS/MIMS reports. This work is informing GSC consensus-building activities and has led to agreement to adopt the Habitat-Lite terminology for use in the Genomic Contextual Data Markup Language (GCDML) (Kottmann et al. 2008).

[1] http://gensc.org/gc_wiki/index.php/MIGS/MIMS

[2] http://environmentontology.org – see the GSCEnvO wiki page for ongoing discussion: http://gensc.org/gc_wiki/index.php/EnvO_Project; also see the EnvO sourceforge site: http://obo.cvs.sourceforge.net/obo/obo/ontology/environmental/.

There is a strong need for developing methods to facilitate the capture of metadata describing the growing number of genomic and metagenomic projects, including information about isolation source and habitat (Morrison et al. 2006; Field et al. 2008). The increase in the associated literature is also accelerating, particularly in light of projects such as the Global Ocean Survey (Venter et al. 2004) and the Human MicroBiome[3] (Gill et al. 2006), with parallel growth in the relevant databases.[4] However, the capture of the metadata associated with these projects remains a major challenge, largely due to the fact that the literature is scattered and the metadata is difficult to find, even by expert manual extraction. Many databases have fields to support the capture of metadata, but such entries are often sparse and are entered as free text, thus lacking standardization in vocabulary and definitions, impeding our ability to perform meaningful comparisons or utilize information from multiple resources. The case studies discussed below illustrate the resulting difficulty in using computational techniques to study the relation between habitat and genotypic or phenotypic properties of organisms (Hunter 2002, von Mering 2006) – a key goal of genomic and metagenomic studies.

Our initial work has focused on a specific metadata type, namely habitat. For our purposes here, we define habitat as "the place or environment where an organism naturally or normally lives and grows." It is distinguished from "sample source", which is the environmental context in which a sample is collected, as defined in Morrison et al.

[3] http://nihroadmap.nih.gov/hmp/
[4] See for example Fig. 1 of (Morrison et al. 2006) for an illustration of exponential growth in the number of sequences in the International Nucleotide Sequences Database Collaboration (INSDC).

2006. Multiple habitat terms can be associated with a species; by contrast, a sample is associated with a description of its (unique) source. Table 1 shows excerpts from the GOLD database (Liolios et al 2008); we can see that the "Habitat" field often has multiple entries, in contrast to the "Isolation" field, which describes the specific sample source and is much more detailed. The initial version of Habitat-Lite is aimed at capturing high level habitat descriptions; ongoing work on the environmental ontology EnvO will provide a much finer grained set of terms to describe specific environments and sample source information.

**** Insert Table 1 here *****

Table 1: Habitat and Isolation fields From the GOLD Database

The development of Habitat-Lite began with the selection of a small list of widely used high-level terms for describing habitat. We used these terms to "bin" information contained in free text fields for habitat or source information in several key databases. This process enables us to develop measures of coverage, utility and usability for the term set, e.g., how well the controlled vocabulary covers the free text entries, how evenly the entries are distributed across the bins defined by the controlled vocabulary, how well the bins capture useful categories for search, how cost-effectively the controlled vocabulary terms can be used to annotate new data, and how consistent the mappings are across multiple annotators (human or automated). There are trade-offs in this complex space between the detailed information that can be captured with a large well-structured

set of terms (e.g., an ontology), vs. the time it takes to create a stable set of structures and the cost of acquiring consistent annotation using this much richer terminology, including supporting tools.

The two major data sources chosen for this study contain large numbers of records and descriptors of habitat in free text form. Ideally, we would have looked in the literature to determine how habitat and isolation source were described. However, for the initial experiments, it was much more efficient to look at fields in existing databases. The two sources were:

1) GenBank[5]: the *isolation_source* field, which captures free text descriptions, in the form entered by submitters to GenBank, related to sample source;
2) Genomes On-Line Database (GOLD)[6] (Liolios et al 2008): the *Habitat* field.which captures terms collected from the literature.

Development of Habitat-Lite

As a starting point, one author (DF) did a survey for terms used in a number of relevant sources. From this list, she selected a set of high level terms as a strawman for the first iteration of the Habitat-Lite term list (shown in Table 2). The number of terms was kept

[5] http://www.ncbi.nlm.nih.gov/Genbank/
[6] http://www.genomesonline.org/

small (less than twenty), based on discussions with annotators at NCBI[7] , but could grow in future iterations. Our approach was to identify a set of seed terms, run experiments to determine how well these could "bin" existing entries, determine how useable such a set of terms would be for human and semi-automated annotation, and then iterate, with the goal of producing a consensus-driven 'minimal set' of habitat terms that provided good coverage of entries in key resources. The original version of Habitat-Lite is available in .obo format[8] and, for example, could be used with OBO Edit[9], CoBrA[10], or the Phenote annotation tool[11].

**** Insert Table 2 here ****

Table 2: Initial Habitat-Lite terms and mapping to EnvO (Oct. 2007)

The initial list of terms drew on previously published lists of habitat terms used to annotate databases (NCBI Microbial genomes[12]), on proposed new community standards for the annotation of 16S sequences[13], on the habitat terms published in the Global Ocean Survey (Nealson and Venter 2007), on habitat terms used to describe the biases in culture collection strains (Floyd et al. 2005) and on patterns and biases in the complete genome collection (Martiny et al. 2005); see Supplementary Table 1 for a full

[7] We met with Tatiana Tatusova, Scott Federhen, Karen Clark and Anji Johnston at NCBI Entrez Genomes; to explore ways to improve the capture of environmental/habitat metadata in GenBank.
[8] http://gensc.org/gc_wiki/index.php/Habitat-Lite
[9] http://oboedit.org/
[10] http://cobra.umbc.edu/eclipse/
[11] Available at http://www.phenote.org/.
[12] http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
[13] http://www.jgi.doe.gov/16s/saiform.php

listing. These terms were mapped to an early version of the Environment Ontology,[14] as shown in column 3 of Table 2.

Use Cases: Analyses Based on Habitat Data

Habitat-Lite terms were assembled from existing terminologies with the explicit goal of supporting as many use cases as possible, in particular, the ability to "bin" data into interesting categories for purposes of comparison. The use of bins is particularly attractive to biologists, who, for example, wish to extract sequences only associated with 'soil bacteria' or 'freshwater metagenomes'. In this respect, biologists' descriptions of 'habitat' contrast strongly with those of environmental scientists who tend to describe habitat in terms of continuous variables.

We are now in the process of assembling use cases to test the coverage of Habitat-Lite. At the 5th GSC meeting, one author (JC) presented a small study done on Ribosomal Database Project (RDP[15], Cole et al. 2007). The RDP consumes GenBank documents for 16S sequences and maintains them in a highly added-value format. These data are used extensively for contrastive analysis based on environmental factors. To determine both the coverage of environments and the utility of the habitat or environmental information in RDP, a small experiment was carried out in late 2006.

[14] http://environmentontology.org
[15] http://rdp.cme.msu.edu/

Using information from the INSDC records, one author (JC) attempted to manually classify into habitats the 168,911 rRNA sequences marked as *environmental* in RDP release 9.44 (November 2006). The habitat categories that were suggested by Phil Hugenholtz (DOE Joint Genome Institute) were modified by splitting *host-associated* into separate categories for *plant* and *animal* (including *human*) *associated*. We first assigned 24.5% of the sequences using their *isolation_source* qualifier. For those sequences without an isolation source tag, or where we were unable to classify based on that tag, we examined the reference titles from the INSDC records and were able to classify another 37.5% of the records. References used by fewer than 150 sequences were not examined because of the effort involved. The remaining 38% of sequences could not be classified because, for the most part, they did not have any habitat information in the INSDC record. Most assignments were made after examination by a single researcher, but spot-checking by a second researcher gave disagreement in assignment for only a small percentage of sequences. By far the biggest category was *animal associated*, and a large fraction of these were *human associated*. The *soil*, *sediment*, and *water* categories also represented large numbers of sequences (see Figure 1).

*** Insert Figure 1 here ***

Figure 1: Categorization of isolation source for environmental sequences from RDP

A second interesting use case was reported in (von Mering et al. 2007). In this paper, the authors studied the association of preferred habitats for microbial clades and looked for correlations between evolutionary distance and similarity of habitat. The habitat

information was taken from free text fields of the Greengenes database (Desantis et al. 2006) and the microbial culture collections (Dawyndt et al. 2005). To assess similarity of habitat, the authors manually selected “informative” words found in the annotation of five or more experiments (von Mering et al. 2007, Tables S2 and S3, supplemental materials) and a computed pairwise similarity score between habitats, based on number of shared keywords. Graphs in Table S2 Figure 2B and 2C show that more habitat “features” are shared among the more closely related organisms, both in terms of taxonomy and molecular similarity.

These use cases illustrate the kinds of information that would be useful to researchers, and also the difficulties of obtaining the information in the absence of a common underlying controlled vocabulary.

GenBank “isolation_source” Entries

To validate and refine the selection of Habitat-Lite terms, examples of habitat or isolation source information were needed to determine what information was present, and how this information was expressed. An ideal approach would be to extract metadata from the published literature, however this is quite difficult, because the metadata occurs in many diverse forms, including PDF tables, densely written materials and methods sections, supplementary material, and even in referenced work. Therefore, we took advantage of the large quantity of free text metadata already available – as fields in database records. As a first step, we analyzed the “isolation_source” field from GenBank gene records

which captures, as short free text entries, information about isolation source of the specific sequence being deposited. John Wilbur (NCBI) provided us with a list of 35,000 distinct isolation_source entries from GenBank gene records as of September 2007 – see Table 3 for examples of some entries from this field.[16]

**** Insert Table 3 here ****

Table 3: Distribution of Unique Habitat-Lite Terms in GenBank *isolation_source* fields

Because of the size of the data set, it was not possible to explore it manually. One of the authors (CC) developed a small set of scripts to identify probable classes based on the presence of specific key words in each entry. The key words used for this analysis were based on the original Habitat-Lite terms plus synonyms and, in some cases, specializations. For example, for "waste water" the terms used for matching were "waste water", "waste-water", "wastewater", "sewage", "sewerage", etc. For "food", the terms used for matching included specific kinds of foods, e.g., "milk.", "cheese," "beer" etc. Similarly, for "organism-associated", the terms used for matching had to capture the many ways of expressing specific organisms, particularly humans, e.g., "M", "patient", "female", "subject", "child" etc.

Of the almost 35,000 distinct entries in the isolation_source field, some 22,000 (63%) contained specific words or phrases that could be mapped to the 17 Habitat-Lite

[16] We were primarily interested in whole genome or metagenomics sequences, but the initial data set consisted of entries for all genes. As a result, frequency counts were heavily skewed towards large metagenomics projects, so we did not use the frequency counts in our analysis. For example, the phrase "locations in the Sargasso Sea, Panama Canal, and the Galapagos Islands" occurred over 3 million times in this data set.

categories. The bulk of these fell into the Organism-associated category (42%). In addition, we were able to identify over 20% of the entries that were geographic names or temporal expressions or other numerical quantities or identifiers. This enabled us to account for approximately 85% of the entries from *isolation_source*. The remaining 15% contain low frequency terms – many of them with species information ("wild mulberry"), location information ("Wilson and the Australian Museum)", or information about culture techniques ("top band of HTA gel").

This pattern-matching approach allowed us to obtain a quick overview of the types of information found in the GenBank isolation_source field. This approach would require significant refinement and/or human intervention if we wished to use it for semi-automated assignment of Habitat-Lite terms to isolation_source entries, for improved search and indexing. In particular, this strategy mapped each entry to a single field, so that, e.g., *130 m below sea surface* was mapped only to "marine," losing the depth information. Similarly, the entry "Marine Biology Laboratory" caused the entry to be associated with the category "Marine" – a plausible inference but certainly not explicit information about habitat.

Habitat Field Entries from the GOLD Database

We next investigated a second data set, which consisted of the Habitat entries from the GOLD database on October 2007. The initial data set consisted of 1455 entries with 2210 terms. Table 1 shows some example GOLD entries, including not only the Habitat field,

but also the much more detailed Isolation field. The entries in the Habitat field frequently contained multiple entries that specified the range of known habitats for a specific organism, e.g., “Host, TB epidemic” or “Aquatic, Soil, Permafrost”.

Coverage of GOLD terms using Habitat-Lite

First, we looked for exact matches between GOLD Habitat terms and Habitat-Lite terms plus the additional term “aquatic”. This resulted in exact matches for 84% of GOLD Habitat terms. The three most frequent terms (“host”, “aquatic” and “soil”) covered 75% of GOLD habitat data, while six Habitat-Lite terms were not seen at all in this smaller data set (“air”, “freshwater”, “extreme”, “microbial mat”, “fossil”, “terrestrial”).

Comparison of automated mapping and expert mapping

In the next experiment, we applied the automated mapping used in the GenBank experiment to the unique entries in the GOLD Habitat data, and compared these results to an expert mapping done by one of the authors (DF). There were a total of 132 unique entries in the GOLD Habitat field for metagenomes. There was 64 % agreement (84/132) and 48 cases of differences in the automated mapping vs. expert mapping. Most differences were due to a failure in the automated mapping procedure (30 cases, which were not classified or not mapped to the limited controlled vocabulary). Another 9 were due to mismatches related to the new category “aquatic” introduced by the expert (5) and 4 were due to difficulty in classifying between freshwater and water.

The remaining nine discrepancies (shown in Table 4) brought to light interesting problems. Several of the discrepancies pointed out an ambiguity in the classification scheme with respect to "extreme environment": terms such as "hot springs", "permafrost", and "hypersaline mats" could be classified as "extreme environment" or into a geographic or environmental feature ("hot springs", "soil", "microbial mat"). In another case ("rice paddies"), it is unclear without further context whether the focus was on the *rice* in rice paddies ("organism-associated") or on the *paddies* ("terrestrial").

*** Insert Table 4 here ****

Table 4: Examples of Disagreement Between Expert Mapping and Automated Mapping for the GOLD Habitat Data.

These examples illustrate well the need for annotation guidelines, to handle situations where a term might be placed in several categories. There are several possible solutions: either there need to be "orthogonal dimensions" that would allow a category like "extreme environment" to be "checked off" separately from some more specific information about geographic or environmental features. Or alternatively, there could be a facility to allow a given term to belong to multiple "bins".

Manual annotation of the GOLD data to two orthogonal bins

The final set of experiments was designed to test the difficulty of the annotation task and to determine whether better annotation could be done by assigning multiple orthogonal

terms. As noted above, there is an advantage to capturing orthogonal annotations, to preserve richer information for searching, and also to reduce interannotator disagreement. To experiment with this approach, a single author (KBC) annotated the 132 GOLD unique terms using Habitat-Lite in conjunction with an explicit set of guidelines that were meant to ensure that every *Habitat* entry was assigned both a general (*biome*) term and an *environment* term. The guidelines made use of the mappings of the Habitat-Lite terms to the EnvO taxonomy as follows:

1. Assign a child term of *biome* (*freshwater, marine,* or *terrestrial*).
2. Can the input be assigned a child class of *habitat* (*organism-associated* or *extreme*)? If so, assign it, and then stop. (This had an undesired effect, which we describe below.)
3. Is the input a food? If so, assign *food.* If not, go to (4).
4. Can the input be assigned a child of *biotic/abiotic* (*biofilms, microbial mat,* or *fossil*)? If so, assign it, and then stop. If not, go to (5).
5. Can the input be assigned a child class of *hydrographical/physiographic/anthropogenic* (*hot spring, hydrothermal vent,* or *wastewater*)? If so, assign it, and then stop. If not, go to (6).
6. Can the input be assigned a child of *environmental substance* (*soil, water, sediment, sludge,* or *air*)? If so, then assign it.
7. Stop.

The undesired effect of Step (2) was that some inputs that could have been assigned specific terms related to extreme habitats were instead only assigned the more general *extreme (habitat)*. A simple re-ordering of the rule might fix this.

The results demonstrate that the annotation task is well within the range of someone with reasonable background in biology. Only 2 out of 132 entries were left unannotated due to lack of domain knowledge: *solfataric fields*, and *self-heated organic materials*. It took approximately 1.5 hours to do about 2 * 132 annotations, or around 1.5 terms annotated/minute. Based on this estimate, it would take less than a day's work to map all of the GOLD Habitat entries to Habitat-Lite.

Discussion

The goals of this work were to create a useful set of high-level terms to capture habitat data, and to develop a methodology that can be applied to similar problems, specifically to:

1) Determine what descriptors of habitat are recorded and how they are expressed in free text;
2) Determine how well a small set of terms, such as Habitat-Lite, could cover terms found in key resources;

3) Examine the feasibility of (semi-)automated capture of the these fields of information for future projects

Our initial experiments have resulted in a new version of Habitat-Lite (shown in Table 5), based on analysis of the GenBank *isolation_source* field and the *habitat* field in the GOLD database. Based on this analysis, we put forward the following recommendations for Habitat-Lite:

- A shift from a 'flat' list to one with some structure is necessary.
- The set of terms should support certain inferences useful for search; for example, that a sample labeled *soil* is also *terrestrial,* or that a sample from a *hydrothermal vent* is also *extreme.*
- Consistent annotation requires guidelines for general terms such as terrestrial and *aquatic,* to instruct annotators to annotate to the most specific term possible.
- The notion of *extreme environment* is problematic in that it should be annotated **in addition** to a more specific term, such as *hot spring* – thus requiring that certain entries be associated with two Habitat-Lite terms.
- The category *Organism-associated* needs to be sub-divided by linking out to other ontologies or controlled vocabularies (specifically, a taxon hierarchy and perhaps a high level anatomy ontology).

- *Fossil* is an example of a currently infrequently used term, but a candidate for inclusion as a term of "exceptional importance" that could be useful in the future for searching.

**** Insert Table 5 here ***

Table 5: Proposed Habitat-Lite Version 2

The new set of Habitat-Lite terms is structured into two levels: a set of high level terms (first column in the table: *aquatic*, *terrestrial*, *air*, plus *organism-associated, food, extreme environment*), and a second level of more specific terms (column 2 in Table 5). To maximize capture of information, this version encourages selection of one or more of the high level terms, one or more of the second level terms, and recording of the specific information in free text (column 3, Table 5). The free text is shown associated with its level 2 term and in column 4, one or more appropriate top level terms.

To maintain simplicity, there is no obligatory connection/restriction between choice of top level terms and second level terms, except for the "food" and "organism-associated" classes. This allows flexibility (for example, there are both freshwater and salt marshes) with the downside of increased possibility for error or for incomplete annotation. It should be possible to do automated association of high level terms, based on the second level terms, e.g., associating "terrestrial" automatically with any annotation of "soil", "sediment," or possible new terms such as "sand", "wood", "rock" or "mud".

The "organism-associated" class should be elaborated by a term describing the organism and an anatomy term for the part of the organism; we will investigate use of a minimal anatomy ontology, such as Jonathan Bard's MIAA (Minimal Information about Anatomy)[17]. The food class for now is just left as free text; it may be possible to use a small specialized food controlled vocabulary or ontology in the future.[18]

Next Steps for Habitat-Lite: Adoption by GOLD, RDP, GCDML

The new version of Habitat-Lite will be tested against the GOLD data and revised to support GOLD (Liolios et al. 2008), IMG[19] (Markowitz et al. 2008) and IMG/M (Markowitz et al. 2008a). GOLD has embraced the adoption of this controlled vocabulary/ontology for its habitat data. Capture of GOLD and IMG habitat data is currently implemented via the Expert Review web submission form on the Integrated Microbial Genomes (IMG) web site. All genomes submitted directly into IMG and IMG/M are now required to provide metadata that conforms to the GOLD vocabulary. The RDP (Cole et al. 2007) has also agreed to adopt the revised version of Habitat-Lite. The new version of Habitat-Lite will be supported in GCDML (Kottmann et al., 2008).

Conclusions

[17] Personal communication.
[18] See http://gensc.org/gc_wiki/index.php/Food_Ontology_Project for discussions about the creation of a food ontology or controlled vocabulary.
[19] For IMG, see http:// img.jgi.doe.gov; For IMG/M, see http://imgweb.jgi-psf.org/cgi-bin/m/main.cgi.

These results indicate that it should be possible to produce a list of terms with good high-level coverage for Habitat-Lite. We accept that candidate Habitat-Lite terms provide only very high-level information and that these terms may be an amalgamation of terms found in different branches of a future ontologies, or even among different orthogonal ontologies (e.g., for “organism-associated”). We also recognize that while these terms may provide a useful tool for biologists and databases, they have severe limitations. We emphasize the importance of maximum reporting of information about habitat, in particular, the necessity of preserving free text fields associated with legacy data so that more fine-grained information is never lost, and re-analysis is always possible.

Long term, our goal is the creation of an interactive metadata checking system (a kind of metadata “spell checker”) that could “read” free text and suggest the correct mapping into a controlled vocabulary/ontology, for user validation or correction, thus ensuring that metadata is comprehensively captured and “binned” at the point of entry.

The use of a combination of Habitat-Lite terms in the short-term, cultural shifts in the way this community annotates to capture more complete descriptions of habitat and isolation source, and future use of ontologies and ontology-aware software will have a measurable benefit on the ability of researchers to effectively re-use ever-growing sources of data for large-scale, downstream analyses.

Towards a minimum information list of habitat terms for use in the GSC

We have posted the initial and revised versions of Habitat-Lite (Table 5) to the GSC wiki. This list is annotated with recommendations and issues which will be addressed in revising this list. We are making an open call for evaluation of this list of habitat terms in order to develop a consensus-driven version of this list that best suits community needs. This terms list will then be implemented in GCDML (Kottmann et al. 2008) and used in the first instance to fill the "Habitat" field of the MIGS compliant Genome Catalogue database (http://gensc.org).

Acknowledgements

The work at MITRE (LH, CC, KBC, SM, JL) has been supported in part by National Science Foundation Grant 0746650, Small Grant for Exploratory Research: Mining Metadata for Metagenomics. We thank Tatiana Tatusova for a number of discussions on an approach to developing a light-weight set of classes for annotation, as well as Scott Federhen, Karen Clark and Anji Johnston of NCBI. We thank John Wilbur, NCBI, for providing us with the data from the GenBank isolation_source fields. We thank Norman Morrison and Lynn Schriml of the EnvO/Gaz project for critical reads of the manuscript.

References

COLE, J.R., CHAI, B., FARRIS, R.J., et al. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Research **35** (Database issue), D169-D172.

DAWYNDT, P., VANCANNEYT, M., DEMEYER, H., SWINGS, J.(2005) Knowledge Accumulation and Resolution of Data Inconsistencies during the Integration of Microbial Information Sources, IEEE Transactions on Knowledge and Data Engineering **17**(8) 1111-1126.

DESANTIS, T. Z., HUGENHOTLZ, P., LARSEN, N (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB, Appl. Environ. Microbiol. **72**(7), 5069-5072.

FIELD, D., GARRITY, G. M., GRAY, T., MORRISON, N., SELENGUT, J. D., STERK, P., TATUSOV, T., THOMAS, N. & ALLEN, M. J. (2008). The Minimum Information about a Genome Sequence (MIGS) specification. Nature Biotechnology, (in press).

FIELD, D. et al. (2008a) Meeting Report: The 5th Genomic Standards Consortium Workshop, OMICS (this issue).

FLOYD, M.M., TANG, J., KANE, M. et al. (2005) Captured Diversity in a Culture Collection: Case Study of the Geographic and Habitat Distributions of Environmental Isolates Held at the American Type Culture Collection, Appl Environ Microbiol. **71**(6): 2813–2823.

GILL, S.R., POP, M., DEBOY, R.T., et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science **312**, 1355-1359.

HUNTER, L. (2002) Ontologies for programs, not people. Genome Biology 3(6).

KOTTMANN, R., GRAY, T., MURPHY, S. et al. (2008) A standard MIGS/MIMS compliant XML Schema: Towards the development of the Genomic Contextual Data Markup Language (GCDML). OMICS (this issue).

LIOLIOS, K., MAVROMMATIS, K., TAVERNARAKIS, N., KYRPIDES, N.C. (2008) The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata Nucleic Acids Res 36, D475-9.

MARKOWITZ, VM., SZETO, E., PALANIAPPAN, K., GRECHKIN, Y., CHU, K., CHEN, I-MA., DUBCHAK, I., ANDERSON, I., LYKIDIS, A., MAVROMATIS, K., IVANOVA, NN., KYRPIDES, N.C. (2008) The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res 36, D528-33.

MARKOWITZ, VM., IVANOVA, NN., SZETO, E., PALANIAPPAN, K., CHU, K., DALEVI, D., CHEN, I-MA., GRECHKIN, Y., DUBCHAK, I., ANDERSON, I., LYKIDIS, A., MAVROMATIS, K., HUGENHOLTZ, P., KYRPIDES, N.C. (2008a) IMG/M: A data management and analysis system for metagenomes. Nucleic Acids Res 36, D534-8.

MARTINY J.B, FIELD, D (2005) Ecological perspectives on the sequenced genome collection. Ecology Letters **8**: 1334–1345.

MORRISON, N., WOOD, J.A., HANCOCK, D., et al. (2006) Standard Annotation of Environmental OMICS Data: Application to the Transcriptomics Domain. OMICS: A Journal of Integrative Biology. **10**(2): 172-178.

NEALSON, K.H., VENTER, J.C. (2007) Metagenomics and the Global Ocean Survey: what's in it for us, and why should we care? The ISME Journal **1**:185-190.

VENTER, J.C., REMINGTON, K., HEIDELBERG, J.F., et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**(5667), 66-74.

VON MERING, C. HUGENHOLTZ, P., RAES, J. et al. (2007) Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments Science **315** (5815), 1126

Organism	STRAIN	PHENOTYPE	HABITAT	ISOLATION
Haemophilus influenzae NTHi	PittEE	Pathogen, Facultative, Nonmotile, Rod-shaped	Host	Middle-ear effusion of a child in Pittsburgh
Mycobacterium tuberculosis	H37Ra	Pathogen, Aerobe, Chemoorganotroph, Rod-shaped, Nonmotile	Host, TB epidemic	Original human-lung H37 isolate in 1934
Psychrobacter sp.	PRwf-1	Psychrophile, Radiation resistant, Rod-shaped, Nonmotile	Aquatic, Soil, Permafrost	
Roseiflexus sp	RS-1	Filament-shaped, Photosynthetic, Thermophile, Facultative, Nonsporulating, Motile, Rod-shaped	Aquatic, Hot spring	Hot spring microbial mat
Lactobacillus reuteri	F275 (JCM 1112)	Probiotic, Non-Pathogen, Rod-shaped, Facultative, Nonmotile	Intestinal flora	Human isolate that is unable to colonize the intestinal tract of mice
Pseudomonas putida	F1	Aerobe, Motile, Rod-shaped, Non-Pathogen	Soil	Polluted creek in Urbana, Illinois by enrichment culture with ethylbenzyne as a sole source of carbon and energy

Table 1: Habitat and Isolation fields From the GOLD Database

Habitat-Lite Terms for genomes and metagenomes		
1	freshwater	ENVO:00000873
2	marine	ENVO:00000447
3	terrestrial	ENVO:00000446
4	soil	ENVO:00001998
5	water	ENVO:00002006
6	air	ENVO:00002005
7	sediment	ENVO:00002007
8	sludge	ENVO:00002044
9	waste water	ENVO:00002007
10	hot spring	ENVO:00000051
11	hydrothermal vent	ENVO:00000215
12	organism-associated	ENVO:00002032
13	extreme environment	ENVO:00002020
14	food	ENVO:00002002
15	biofilm	ENVO:00002034
16	microbial mat	ENVO:01000008
17	fossil	ENVO:00002164

Table 2: Initial Habitat-Lite Terms and Mappings to EnvO (Oct. 2007)

Class	Table Frequency	Data Set Frequency	Percent Total	Example
ORGANISM_ASSOCIATED	14781	341003	42.4%	1 year old male spleen
WATER/AQUATIC	2008	40794	5.8%	0 m water at a station in the North Atlantic
SOIL	1115	229032	3.2%	0-20 cm bulk soil from a mixed forest
MARINE	944	3115879	2.7%	0.2-0.8 um fraction from surface sea water
SEDIMENT	723	34435	2.1%	aquaculture coastal sediments
TERRESTRIAL	595	3100550	1.7%	a declining forest
FOOD	398	4003	1.1%	(onion)
SLUDGE	294	9868	0.8%	1st maturation stage of sludge
MICROBIAL MAT	195	9164	0.6%	a deep sea microbial mat
WASTE WATER	195	5969	0.6%	activated tannery effluent from treatment plant
HYDROTHERMAL VENT	133	3036	0.4%	14 N Mid Atlantic Ridge Logatchev vent field
HOT SPRING	121	3249	0.3%	6-48 celsius region of a hot spring
EXTREME	117	4967	0.3%	a solar saltern
BIOFILM	114	3499	0.3%	aquatic phototrophic biofilm
FRESHWATER	75	2609	0.2%	Arctic freshwater lake
FOSSIL	67	507	0.2%	100,000 year old fossil
AIR	21	768	0.1%	African air sample
TOTAL HABITAT-LITE TERMS	21896		62.9%	
TOTAL UNIQUE	34836			

Table 3: Distribution of Unique Habitat-Lite Terms in GenBank *isolation_source* Fields

GOLD HABITAT Term	Expert Mapping	Automated Mapping
Mud	terrestrial	Soil
Rice paddies	terrestrial	Organism-associated
Soda lakes	organism-associated	Water
Hot Spring	extreme environment	Hot spring
Hot spring	extreme environment	Hot spring
Permafrost	extreme environment	Soil
Snow	extreme environment	Freshwater
Sulfur spring	extreme environment	Water
Hypersaline mats	microbial mat	Extreme

Table 4: Examples of Disagreement Between Expert Mapping and Automated Mapping for the GOLD Habitat Data.

TOP LEVEL	Second Level:	Third Level:	Example also could be coded for:
Choose one or more:	**Choose one or more:**	**Free text description, e.g.,**	
Aquatic: freshwater	soil	pinyon-juniper forest soil	Terrestrial
Aquatic: marine	sediment	oxygen-depleted intertidal marine sediment	Aquatic: marine
Aquatic	sludge	thermophilic methanogenic sludge	Terrestrial?
Terrestrial	waste water	waste water of paper machine	Aquatic
Air	hot spring	hot spring at 70 degrees C	Aquatic, Extreme
Fossil	hydrothermal vent	the shallow hot vent in IwoJima	Aquatic: marine, Extreme
	biofilm	biofilm of drinking water distribution system	Aquatic
	microbial mat	hot spring microbial mat	Aquatic, hot spring
Food	[Food Ontology or CV]	surface of smear ripened cheese	Food
Organism Associated	[Species CV] [Anatomy CV, e.g., MIAA]	gut of nitidulid beetle	Organism-Associated
Extreme Environment	Select if appropriate	extremely alkaline (ph 12 to 13) groundwater	Aquatic; Extreme
Other		(45.32739 N , 80.40874 W)	

Table 5: Proposed Habitat-Lite version 0.2